Volume 12 • Number 1 • 2009
VALUE IN HEALTH

Reliability of the Discrete Choice Experiment at the Input and Output Level in Patients with Rheumatoid Arthritis

Ulla Slothuus Skjoldborg, PhD,[1] Jørgen Lauridsen, PhD,[2] Peter Junker, MD[3]

[1]Unaffiliated; [2]Institute of Public Health, University of Southern Denmark, Odense, Denmark; [3]Odense University Hospital, Odense, Denmark

ABSTRACT

Objectives: To investigate the issue of conjoint reliability over time.
Methods: A discrete choice experiment was applied using scenarios that describe the effect of treating rheumatoid arthritis patients with TNF-alpha inhibitors, a novel class of highly effective, but expensive antirheumatic agents. Respondents participated in three face-to-face interviews over a period of 4 months. Reliability was measured both at the input level, where the consistency of matches made by respondents to the Discrete Choice Experiment (DCE) question between replications was determined, and at the output level, where the parameters of the conjoint model were estimated and tested for joint significance and willingness to pay (WTP) confidence intervals were calculated.
Results: Input level: Of the 1661 choices made in survey 1, 1316 were repeated in survey 2. Based on the observed number of consistently repeated choices and the expected number by chance, a fair agreement between the choices in the two surveys ($\chi^2 = 324$) was found. Of the 998 consistently repeated choices from survey 1 to survey 2, 818 were repeated in survey 3. There was again a high level of consistency between the choices in surveys 1 and 2 and the final choice in survey 3. Output level: The confidence intervals for WTP figures in surveys 1 and 2 and 1 and 3 were overlapping, implying that the DCE was reliable at the output level over time.
Conclusion: The proportion of consistent responses was higher than would be expected by chance. Conjoint reliability over time was found both at the input and output level.
Keywords: conjoint measurement, discrete choice experiment, reliability, rheumatoid arthritis, TNF-alpha inhibitor treatment.

Introduction

Conjoint analysis (CA) is a method that can be applied to assess preferences and willingness to pay (WTP) [1–3]. Because CA is a more and more popular method of choice when investigating preferences, it is of high importance to explore the methodological problems that might be related to the method. Among others, studies have considered methodological issues, such as inconsistency, heterogeneous preferences, and dominating attributes. Another very interesting and important factor to explore is reliability over time, which will be the objective of this article.

CA is based on random utility theory, where the utility of the good being evaluated is considered as being composed of a deterministic part, which is interpreted as an indirect utility function, and a random part, which is assumed to be composed of factors that are not observable, but influence utility. This random element measures errors in the dependent variable and/or model specification errors.

When applying CA, hypothetical scenarios are presented to the respondent, each describing different levels of the attributes that characterize the good being evaluated. The respondents' preferences are measured by asking them to state which alternative they prefer. It is possible to determine the relative importance of the attributes; i.e., the marginal rate of substitution when giving the attributes different values. The total explained utility for different combinations of attributes can furthermore be estimated, thereby identifying the respondents' most favored preferences. WTP estimates can be calculated by including a cost attribute.

Several aspects of reliability could be considered [4,5]. In the present investigation, we focused on temporal reliability, implying that measurements were repeated using the same instrument and the same respondents at least two times with separate time intervals. Using this test–retest method, a coefficient r (also known as the reliability estimate), which measures the correlation between the sets of observations, is determined as a ratio of the true variance ($S^2_{\text{true observation}}$) to the observed variance of observations ($S^2_{\text{observed observation}}$) between measurement periods x_1 and x_2

$$r_{x_1, x_2} = S^2_{\text{true observation}} / S^2_{\text{observed observation}}$$

To the authors' knowledge, this approach has only been used once before in a health-care setting, by Bryan et al. [6]. The present study, however, differs from the previous report by using a much longer time interval between measurements. By extending the time between measurements, we try to minimize the pitfall of having the respondents repeatedly answering the same questions (later statements being influenced by previous made statements), a limitation of the test–retest method which could cause problems for a study like Bryan et al.'s were only short periods (approximately 2 weeks was stated as the longest period of time between questionnaires) were applied. Furthermore, the present study differs by applying a set of hypothetical alternatives that appear more complex than in the study by Bryan et al., hereby exploring the reliability of CA further by using a more challenging setting which in addition might be more realistic in an environment like health care, where choices are complex and difficult to make. Finally, as explained further in the next section, we investigate output reliability, a subject which is not considered by Bryan et al. and which has only been explored in a few stated preference surveys, mostly in studies using Contingent Valuation (CV) and not Stated Choice (SC) studies.

The strategy of the article was as follows. First, we considered reliability from an input variable point of view, considering how consistent the respondents' answers to the CA-questions were over time. Next, reliability was investigated by looking at the outcome of the CA analysis; i.e., we determined the consistency of the different attribute weights over time. Finally, we specified WTP to vary by survey, as in Bryan et al. [6]. Output reliability is an issue of interest to CA, as surveys using stated preferences

Address correspondence to: Ulla Slothuus Skjoldborg, Amerika Plads 8, 4.tv. DK-2100 Copenhagen, Denmark. E-mail: ulla@skjoldborg.biz
10.1111/j.1524-4733.2008.00402.x

Table 1 The different attribute values used in the model

Attribute no.	Attributes	Values
1	Duration of morning stiffness	0; 5; 30; 60; 90; 120
2	Pain level	0; 2; 4; 6; 8; 10
3	Number of swollen joints	0; 5; 10; 15; 20; 25
4	Feeling of being tired	Reduced (0); unchanged (1)
5	Slightly higher risk of a minor infection	Yes (1); no (0)
6	Out-of-pocket payment per month in excess of present expenditure for arthritis medication (DKK)	0; 50; 100; 200; 450; 575; 800; 900; 1075; 1150; 1250; 1500; 2150; 2300; 2500; 3000; 4300; 5000

have for the most part not explored the effect of time on the respondents' WTP responses. Only a few CV surveys [7,8] have investigated this aspect, and, with regard to SC experiments, only one study [9] was found to consider this issue. Because of the lack of research on this subject, and because of the fact that SC increasingly appears to be the choice of method applied, we believed output reliability to be an important issue to investigate in relation to the present study. The conjoint measurement scenarios were applied to patients with rheumatoid arthritis (RA), a chronic, potentially debilitating inflammatory joint disease, which affects around 1% of most populations. The study was focused on treatment with TNF-alpha blockers, a new and highly promising treatment option for RA patients, who have not benefited from traditional disease modifying antirheumatic drugs [10]. This novel therapeutic modality is expensive, however, amounting to around US$20,000 per year for each individual patient.

Method and Setting

In this study, a discrete choice format was applied to investigate reliability in a population of rheumatoid arthritis respondents. Reliability was defined as consistency of results; i.e., a reliable measure had no variation in the observed score because of random errors [4].

A total of 325 patients diagnosed with RA—i.e., RA according to the 1987 American College of Rheumatology (ACR) classification criteria—who received therapy at the outpatient clinic at Odense University Hospital, Section of Rheumatology, were asked to participate in the study. The outpatient clinic is a tertiary center, serving the County of Funen. These 325 patients encompassed the total number of RA patients aged between 18 and 70 years, who were registered with the clinic as of July 2003. Patients received a letter of introduction describing the study, and were subsequently contacted by phone regarding participation. One hundred seventy-eight agreed to participate.

Choice of Attributes

The attributes included in the description of the treatment scenarios are presented in Table 1. These attributes were derived from commonly used health status instruments like the Nottingham Health Profile [11] and the American College of Rheumatology definition of improvement in RA (ACR response criteria), a composite measure which includes a selection of clinical and laboratory items reflecting disease activity and patients physical function [12]. These instruments are being widely applied in controlled clinical trials and are considered to be equally relevant to studies on conventional antirheumatic agents and biologics. Concerning the cost attribute, levels from an earlier study investigating RA patients' WTP for RA treatment [13] were applied, because this range showed that the maximum levels of payment were adequately high to ensure that maximum WTP estimates could be obtained.

Attributes listed in Table 1 were included as explanatory variables in a random effect logit model.

Selection of Scenarios

Given the number of attributes and the number of possible outcomes per attribute, the total number of possible combinations was exceedingly high ($2^2 \times 6^3 \times 18 = 15{,}552$), necessitating a systematic reduction in number of scenarios applied while maintaining optimal design criteria [14,15]. This was accomplished by establishing a fractional factorial design, assuming interactions among attributes to be insignificant. An experimental design optimizing D-efficiency was generated in SAS [16]. The procedure in SAS simultaneously chose alternatives and paired the alternatives to choice sets, resulting in a design with 62 choice sets (D-coefficient = 97.82). These choice sets were grouped into eight (as orthogonal as possible) blocks of choices sets with 10 choices each. The survey was thus divided into eight questionnaires, each of which consisted of 10 choice exercises. The eight subgroups of respondents assigned to each questionnaire were tested to be homogenous with respect to age, sex, and respondent's duration of illness [17].

It has previously been shown [18] that respondents are capable of handling up to 13 discrete choice questions per interview. In the present study, each interview comprised 10 such questions.

The Interviews

The respondents participated in three face-to-face interviews 4 months apart. In each of these interviews, they were asked to indicate their priorities among a selection of treatment outcomes. An example of a choice situation is shown below in Table 2. The respondents were asked the following: "The cards A and B each describe outcome and cost of treatment. Which one would you prefer?"

In addition to the discrete choice questions, information about respondents' socioeconomic characteristics was collected in the first interview. These questions were asked at the end of the interview to ensure that there were no differences in the three

Table 2 Example of CA question

Attributes	Card A	Card B
Duration of morning stiffness	90 minutes	60 minutes
Pain level	8	10
Number of swollen joints	20	0
Feeling of being tired	Reduced	Unchanged
Slightly higher risk of a minor infection	No	Yes
Out-of-pocket payment per month in excess of present expenditure for arthritis medication (DKK)	2300	900

CA, conjoint analysis.

Table 3 Variables included in the model

Variable name	Description of variable
Duration of illness	The length of time the respondent has been diagnosed with arthritis (years)
Reported degree of morning stiffness	The respondents' own valuation of experiencing morning stiffness on a scale from 0 to 10 (10 being the worst)
Reported degree of pain	The respondents' own valuation of experienced pain on a scale from 0 to 10 (10 being the worst)
Reported degree of swollen joints	The respondents' own valuation of experiencing swollen joints on a scale from 0 to 10 (10 being the worst)
Reported degree of tiredness	The respondents' own valuation of experiencing tiredness on a scale from 0 to 10 (10 being the worst)
Reported degree of adverse effects	The respondents' own valuation of experiencing adverse effects on a scale from 0 to 10 (10 being the worst)
Prescriptive drug	Does the respondents have a monthly expenditure for prescriptive drugs (yes, no)
TTO	The EQ-5D estimate (Danish weights [8] have been used for calculation purposes)
Birth cohort	The respondents' year of birth
Sex	The respondents' sex (male = 0; female = 1)
Civil status	The respondents' civil status, (single = 0, married/cohab = 1)
Occupation	Self-employed = 1, public/private employed = 2, retired = 3, other nonemployed = 4
Income	The respondents yearly income before tax (in 1000 DKK)

TTO, Time Trade Off.

interviews concerning the timing of the DCE questions. Furthermore, to be able to adjust for potential interference by health status changes between interviews, questions about the patients' clinical condition were included on all three occasions.

Each respondent completed the same block of questions at each administration; the order in which respondents completed the questions was also held constant.

Analytical Model

A linear additive utility function was assumed; i.e., a rise in the value of one attribute would give a proportional rise or fall in total utility. Furthermore, it was assumed that the utility associated with one attribute was not affected by the utility experienced from another attribute. A basic model describing the utility associated with the effect of a given TNF-alpha inhibitor treatment relative to an alternative option was therefore described as:

$$\Delta U = \beta_1 \cdot \Delta x_1 + \beta_2 \cdot \Delta x_2 + \beta_3 \cdot \Delta x_3 + \beta_4 \cdot \Delta x_4 + \beta_5 \cdot \Delta x_5 + \beta_6 \cdot \Delta x_6 + \varepsilon + \mu$$

where six attributes were included as explanatory variables. $\Delta x_1, \ldots, \Delta x_6$ represent the differences in attribute values between alternative A and alternative B, $\beta_1, \ldots \beta_6$ are the attribute specific weights, and ΔU the change in utility as a result of choosing alternative B instead of alternative A. The error term ε is the random error term, including random variation across discrete choices, and μ is the random variation across respondents.

The utility of alternative A was defined to be zero, which implied that $\Delta U > 0$ if B generated higher utility than A, and $\Delta U < 0$ if B generated lower utility. It was assumed that the individual would choose alternative B only if $\Delta U > 0$.

Along the lines of Bryan et al. [6], we allow the WTP estimates to vary by hypothesizing that the attribute weights $\beta_1, \ldots \beta_6$ varied with survey by simply inserting indicator variables S_2 for survey 2 and S_3 for survey 3. In extension of Bryan et al. [6], we furthermore allow the WTP estimates to vary with individual characteristics $z_1, \ldots, z_K$ (including severity of disease and second-order degrees of continuous characteristics). Thus, we specify the attribute weights as

$$\beta_j = \gamma_{1j} + \gamma_{2j} S_2 + \gamma_{3j} S_3 + \alpha_{1j} + \cdots + \alpha_{Kj} z_K, \; j = 1, \ldots, 6$$

Operationally, the γ_{ij} and α_{ij} parameters were estimated using a logistic regression with product variables of the attribute variables versus the dummies for surveys and the individual characteristics. The regression was adjusted for random individual effects to ensure efficient estimation. A few studies [19–21] have adjusted for heteroscedastic scale and found some evidence that this adjustment matters for WTP estimates. Nevertheless, as this adjustment runs maximum likelihood estimation highly complex, and thus increases the risk of unreliable estimates, it was decided to omit it for the present study. Next, the β_j parameters were calculated. Finally, WTP was calculated for each attribute.

Variances were calculated using the Krinsky-Robb method. Ninety-five percent confidence intervals for WTP for attributes 1 to 6 were created as a function of individual characteristics. This was done by letting the individual characteristics vary over the sample range (against the remaining characteristics on a sample average). For each value of the individual characteristic, 10,000 Krinsky-Robb replications were made.

Variables Included in the Model

A few sociodemographic variables, the EQ-5D estimated Time Trade Off (TTO), and variables describing the extent of inconvenience associated with having arthritis, were, in addition to the income variable, included in the model. The variables included in the model are presented in Table 3. These variables were selected, partly based on suggestions in previous studies, and partly based on availability of information from the survey. The choice of the TTO was justified by its operational simplicity, i.e., it includes only a few questions to be asked as opposed to health measures of higher dimensionality. Furthermore, the restricted data availability did not allow for inclusion of data providing more complete descriptions of demographic, disease, or comorbidity. Especially, inclusion of disease history might affect the CA results.

Results

A total of 178 out of 325 candidate respondents participated (55%). Of these, 145 (45%) and 130 (40%) participated in the second and third survey, respectively.

Details about the attributes and their relative weights for the three surveys have been presented previously [17]. As mentioned in the methods section, a few sociodemographic variables, the TTO variable, and variables describing the extent of inconvenience associated with having arthritis were, in addition to the income variable, included in the model. A few variables included were not significant, but were still incorporated in the model, as they appeared to be important from a theoretical standpoint.

Table 4 Tabulation of repeated choices in survey 1 and 2

	Survey 2		
	A	B	Total
Survey 1			
A	632 (476)	154 (311)	786
B	164 (321)	366 (209)	530
Total	796	520	1316

Chi-square = 324.0, P < 0.0001. Numbers in parentheses are expected number by chance (i.e., row total multiplied by column total divided by grand total).

This is the case for the following variables: TTO, prescription drugs, occupation, and income. The majority of variables, however, was significant, and appears to influence the choice of card A or card B.

Reliability at the Input Level

The consistency of matches made by respondents to the DCE question between replications was determined.

In the first survey, the 178 respondents were allowed to make 10 choices, giving a total of 1780 choices. Nevertheless, as some of the choices were refused, a total of 1661 choices were completed. Of these 1661 choices, 1316 were repeated in survey 2. Table 4 presents a tabulation of these. The observed number of consistently repeated choices was $(366 + 632) = 998$, which was equivalent to $(998/1316) \times 100\% = 75.8\%$. The expected number by chance was $(209 + 475) = 684$, which was equivalent to $(684/1316) \times 100\% = 52.0\%$, and thus there was a fair agreement between the choices in the two surveys ($\chi^2 = 324$).

Subsequently, of the 998 consistently repeated choices from survey 1 to survey 2, 818 were repeated in survey 3 (Table 5). The observed number of consistently repeated choices was 713, which was equivalent to 87.2%, while the expected number by chance was 437 or 53.4%, thus indicating a good correspondence between a consistent choice in survey 1/2 and the final choice in survey 3.

Reliability at the Output Level

The parameters of the conjoint model were estimated and tested for joint significance (see Skjoldborg et al. [17] for details). It was found [17] that survey 2 did not differ from survey 1 with regard to the coefficients in the logistic regression. Estimation results for the attributes and their relative weights for the three surveys showed that except for attribute 5 (adverse effects) in survey 2 (which just kept its place on a 10% significant level), the interactions were not significant [17]. Thus, Skjoldborg et al. [17] concluded that survey 2 did not differ from survey 1. A similar conclusion appeared when looking at surveys 1 and 3.

Table 5 Tabulation of repeated choices in survey 1 to 2 (those who were consistent in 1 and 2) and 3

	Survey 3		
	A	B	Total
Survey 1			
A	464 (326)	61 (199)	293
B	44 (182)	249 (111)	525
Total	508	310	818

Chi-square = 430.1; P < 0.0001.
Note: see Table 4.

Nevertheless, because WTP was a nonlinear function of parameters, it was necessary to take a closer look at the confidence intervals of the WTP's before making any final conclusions.

The confidence intervals for WTP figures in surveys 1 and 2 and 1 and 3 were overlapping, implying that DCE was reliable at the output level over time (Table 6).

The presence of dominating attributes, i.e., the phenomenon that individuals consistently choose the alternative with the best value on a certain attribute, was investigated. Table 7 shows the percentage of individuals who did so for each of the six attributes. In particular, pain and payment appeared to be dominating attributes for 12.41% and 9.89%, respectively, of the respondents. Furthermore, Table 7 shows the Spearman rank correlations between individual characteristics and presence of dominance on each of the six attributes. Pain seems especially to be a dominating attribute for respondents with a short duration of illness, the younger birth cohorts, self-employed, and nonretired, while payment seems to be dominating for elder cohorts, those who are not self-employed, the retired, and low-income respondents. Apart from these, risk of minor infection seems to be a dominating attribute for prescriptive drugs receivers. The indicators for surveys are not related to dominance on any attribute, which shows that the dominance pattern is constant across surveys.

The final column of Table 7 shows that the percentage of individuals showing internal inconsistency is low (%1.84) and that inconsistency is especially related to elder cohorts and to the retired. The indicators for surveys are not related to inconsistency, thus showing that the inconsistency pattern is unrelated to survey.

Discussion

In this study involving repeated discrete choice experiments among patients with RA, evidence of reliability was found at both input and output level.

When investigating temporal reliability, the interval between surveys is of crucial importance. Short intervals may imply a memory effect; i.e., the respondents are able to recall their previous answers. With long intervals, on the other hand, preferences may become influenced by confounding factors; e.g., altered health status. In the present study, 4-month intervals were adopted. To be able to adjust for potential interference from health status changes that might have occurred between the interviews, questions about the patients' clinical condition were

Table 6 WTP (1000 DKK) by survey, with Krinsky-Robb 95% CI

Attribute	Survey	WTP	Lower	Upper
Morning stiffness	1	0.00754	0.00380	0.01127
Morning stiffness	2	0.00726	0.00279	0.01176
Morning stiffness	3	0.00398	–0.00086	0.00881
Pain	1	0.22603	0.16495	0.28711
Pain	2	0.22219	0.15960	0.28477
Pain	3	0.23385	0.17047	0.29722
Swollen joints	1	0.02872	0.00757	0.04986
Swollen joints	2	0.01404	–0.00918	0.03727
Swollen joints	3	0.01477	–0.00921	0.03876
Tiredness	1	0.82046	0.47343	1.16749
Tiredness	2	0.54162	0.14309	0.94016
Tiredness	3	0.34620	–0.01981	0.71220
Adverse effects	1	0.69515	0.36738	1.02291
Adverse effects	2	0.41129	0.07792	0.74465
Adverse effects	3	0.49293	0.11934	0.86653

CI, confidence interval; WTP, willingness to pay.

Table 7 Analysis of dominating attributes

	Indicator for dominance of attribute						Inconsistency rate
	Morning stiffness	Pain	Swollen joints	Tired	Risk of infection	Payment	
% respondents showing dominance/inconsistency	3.91	12.41*	0.92	2.99	2.07	9.89	1.84
Pearson rank correlations							
Duration of illness	–0.07	–0.14	–0.01	–0.04	–0.02	0.04	0.08
Reported degree of morning stiffness	0.01	–0.05	–0.04	–0.11	0.01	0.05	0.05
Reported degree of pain	–0.01	0.05	0.04	–0.05	0.06	–0.01	–0.03
Reported degree of swollen joints	–0.05	–0.02	0.10	–0.04	–0.02	0.04	0.02
Reported degree of tiredness	–0.01	–0.08	0.04	0.11	0.02	–0.04	0.08
Reported degree of adverse effects	–0.04	0.10	–0.01	0.02	–0.05	0.02	0.02
Prescriptive drug	–0.05	–0.02	–0.02	0.01	0.15*	0.06	–0.03
TTO	0.11	0.08	–0.04	–0.01	–0.06	–0.05	–0.06
Birth cohort	0.02	0.13*	–0.05	0.04	0.03	–0.19*	–0.14*
Sex	–0.01	–0.05	0.06	0.06	–0.07	0.01	0.02
Civil status	–0.01	–0.04	–0.01	–0.01	0.04	–0.01	–0.04
Self-employed	0.06	0.21*	–0.05	–0.01	–0.05	–0.14*	–0.08
Employed	–0.01	0.06	–0.02	0.01	0.02	–0.08	–0.03
Retired	–0.01	–0.21*	0.08	–0.04	–0.01	0.16*	0.12*
Other nonemployed	–0.06	–0.02	–0.03	0.08	0.06	0.01	–0.04
Income	0.03	0.12	0.02	0.03	–0.02	–0.16*	–0.10
Indicator for survey 1	–0.06	0.01	0.02	–0.02	0.05	–0.02	–0.01
Indicator for survey 2	0.01	–0.02	0.03	–0.01	–0.06	–0.01	0.01
Indicator for survey 3	0.05	0.01	–0.06	0.03	0.01	0.04	–0.01

Significance of Spearman Rank correlation at 1% level marked by *.
TTO, Time Trade Off.

included on all three occasions. In addition, we tested whether the average TTO score remained constant through the three surveys (results not reported). This hypothesis could not be rejected, and hence we assumed that health status was unchanged.

Theoretically, the respondent drop-out between surveys 2 and 3 may have influenced the results. Nevertheless, the dropout rate was small, and baseline data do not indicate that the drop out subset represents a subgroup with distinctive characteristics. We have no specific information about their reasons for nonadherence to the protocol.

Most variables had a significant effect on utility [17]. Nevertheless, a few variables were not significant, but still we chose to include them in the model due to theoretical importance. This is the case for respondent's income. The income variable was of importance, because the respondent's WTP would be expected to be influenced by the variable—that is, if the respondent kept his budget restriction in mind when looking at the cost involved in the choice scenarios. This variable appeared to be significant in a nonadjusted logit model, but came out as insignificant when the model was controlled for random individual effects.

To the authors' knowledge, measuring reliability in the health-care field has only been done once before by Bryan et al. [6]. Bryan et al. also investigated reliability over time. Their results were promising, indicating a high reliability level at both the input data and result levels. The present data accord well with those by Bryan et al. by demonstrating a high level of reliability at both input and output levels. In our study, however, we applied a much longer time interval between measurements and a set of hypothetical alternatives that seemed more complex than in Bryan et al. [6]. In addition, different statistics were applied to study the reliability at the input level. In the present study, chi-square was applied, whereas Bryan et al. used the kappa (κ) statistic. We consider chi-square to be more convenient, because it could be used to derive an explicit probability for classification-by-chance, while the κ statistic is rather an ad hoc measure that is open for interpretation. At the output level, we used the Wald test, because the finite-sample F tests applied by Bryan et al. [6] are not strictly valid in the asymptotically justified maximum likelihood estimation framework, which is applied for the CA. Further, at the result level, we included the models by Bryan et al. to incorporate not only shifts in WTP's across surveys, but also across individual characteristics. This was considered to be important to avoid bias, because if changes in the WTP values over surveys were related to individual characteristics (which they indeed were [17]), models including only survey shifts as those presented by Bryan et al. [6] would be biased.

Looking for presence of dominating attributes, it was found that particularly pain and payment seem to be dominating. As the Spearman rank correlations (Table 7) further explores, pain seems to be a dominating attribute especially for respondents with a short duration of illness, the younger birth cohorts, self-employed, and nonretired. These respondents might feel more limited and affected by the pain they experience, as they presumably have less experience with RA-related pain and have a more active life (they are working and they are younger) than the remaining groups of respondents. Payment seems to be dominating for elder cohorts, those who are not self-employed, the retired, and low-income respondents—overall respondents who might face a more limited budget and hence are more sensitive to expenditures. Risk of minor infection seems to be a dominating attribute for prescriptive drugs receivers. Indicators for surveys were not related to dominance on any attribute, which shows that the dominance pattern is constant across surveys. It is an open question to which extent people dominated on one feature and focused on that throughout, or to which extent their dominance rather expressed consistency. These two sources cannot be exactly separated by the data. Nevertheless, the highest dominating factor was only about 12%. Apart from being reassuring in itself, this figure serves as an upper limit to the extent of expressed dominance.

Study Limitations

The respondents participated in face-to-face interviews. This could be considered problematic from a theoretical point of view because the interviewer could have affected the outcome.

A more potential type of limitation may be that many surveys are now provided to respondents online. Hence, in theory, the results generated from face-to-face interviews could differ from this more objective type of setting.

Even though several attributes are applied in this study, thereby increasing the complexity of choosing among the alternatives, it is possible that more complex surveys (particularly those that employ a Likert scale response format) may result in lower observed reliability than observed in this study.

Conclusion

This study demonstrated conjoint reliability over time in DCE at both the input and the output level. Future work investigating reliability in conjoint measurement using even more complex settings than in this survey, and more research on the effect of time on respondents' stated WTP (output reliability) is encouraged to further explore this important methodological issue.

Source of financial support: Financial support by the Danish Medical Research Council is acknowledged.

References

1 Beggs S, Cardell S, Hausman J. Assessing the potential demand for electric cars. J Econ 1981;16:1–19.
2 Train K. Qualitative Choice Analysis. Theory, Econometrics and an Application to Automobile Demand. Cambridge, MA: MIT Press, 1986.
3 Cramer JS. The Logit Model for Economists. London: Edward Arnold, 1991.
4 Frankfort-Nachmias C, Nachmias D. Research Methods in the Social Sciences (5th ed.). New York: Oxford University Press, 1996.
5 Bateson JE, Reibstein D, Boulding W. Conjoint analysis reliability and validity: a framework for future research. In: Houston MJ, ed., Review of Marketing. Chicago: American Marketing Association, 1987.
6 Bryan S, Gold L, Sheldon R, Buxton M. Preference measurement using conjoint methods: an empirical investigation of reliability. Health Econ 2000;9:385–95.
7 Whittington D, Smith VK, Okorafor A, et al. Giving respondents time to think in contingent valuations studies: a developing country application. J Environ Econ Manage 1992;22:202–25.
8 Lauria DT, Whittington D, Choe K, et al. Household demand for improved sanitation services: a case study of Calamba, Phillipines. In: Willis K, Bateman I, eds., Valuing Environmental Preferences: Theory and Practice of the Contingent Valuation Method. Oxford: Oxford University Press, 1999.
9 Cook J, Whittington D, Canh DG, et al. Typhoid vaccines with time to think in Hue, Vietnam. Econ Inq 2007;45:100–14.
10 Furst DE, Breedveld FC, Kalden JR, et al. Updated consensus statement on biological agents, specifically tumour necrosis factor α (TNFα) blocking agents and interleukin-1 receptor antagonist (IL-1ra) for the treatment of rheumatic diseases. Ann Rheum Dis 2004;63(Suppl. II):S2–12.
11 Kind P, Carr-Hill R. The Nottingham Health Profile: a useful tool for epidemiologists? Soc Sci Med 1987;25:905–10.
12 Felson DE, Anderson JJ, Boors M, et al. American College of Rheumatology preliminary definition of improvement in rheumatoid arthritis. Arthritis Rheum 1995;38:727–35.
13 Slothuus U, Larsen ML, Junker P. Willingness to pay for arthritis symptom alleviation. Comparison of closed-ended questions with and without follow-up. Int J Technol Assess Health Care 2000;16:60–72.
14 Street DJ, Bunch DS, Moore BJ. Optimal designs for 2^k paired comparison experiments. Commun Stat Theor M 2001;30:2149–71.
15 Burgess L, Street DJ. Optimal designs for 2^k. Choice experiments. Commun Stat Theor M 2003;32:2185–206.
16 Kuhfeld WF. Marketing research methods in SAS experimental design, choice, conjoint, and graphical techniques. TS-722. 2005. Available from: http://www.sas.com [Accessed March 4, 2008].
17 Skjoldborg US, Lauridsen J, Junker P. Conjoint reliability investigated at the input and output level. Health Economics Papers 7, Institute of Public Health, SDU 2005. Available from: http://www.sam.sdu.dk/healtheco/publications/20057.pdf [Accessed March 4, 2008].
18 Ryan M, Hughes J. Using conjoint analysis to assess women's preferences for miscarriage management. Health Econ 1997; 6:261–74.
19 Swait J, Adamowicz W. Choice environment, market complexity, and consumer behaviour: a theoretical and empirical approach for incorporating decision complexity into models of consumer choice. Organ Behav Hum Decis Process 2001;86:141–67.
20 Dellaert BGC, Brazell JD, Louviere JJ. The effect of attribute variation on consumer choice consistency. Mark Lett 2006; 10:139–47.
21 Bech M, Gyrd-Hansen D, Kjær T, et al. Graded pairs comparison—does strength of preference matter? Analysis of preferences for specialised nurse home visits for pain management. Health Econ 2007;16:513–29.